

DOCUMENT RESUME

ED 191 877

TM 800 498

AUTHOR Haney, Walt
TITLE Standards for Tests and Test Use. National Consortium on Testing Staff Circular No. 3.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY National Consortium on Testing, Cambridge, Mass.
PUB DATE Oct 78
NOTE 48p.
AVAILABLE FROM National Consortium on Testing, P.O. Box 9521, Arlington, VA 22209 (\$3.00)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Codes of Ethics; Credibility; *Educational History; Educational Testing; Predictive Measurement; Psychological Testing; *Standards; Student Rights; Test Construction; *Test Interpretation; *Test Reliability; *Test Validity

IDENTIFIERS *American Psychological Association; Test Manuals; Test Publishers; Test Security

ABSTRACT

With the intent of spurring discussion and raising important issues about the topic of test standards, this circular serves primarily as a resource document for member organizations of the National Consortium on Testing (NCT). Additionally, it identifies the broad changes which have evolved since 1953 regarding ethical and technical standards for testing. The most apparent change in the two sets of standards is that while the ethical standards have diminished in number, the technical standards have multiplied and their rate of change is accelerating faster than the ethical standards. Specific issues arising from this review include: (1) the emphasis on related validity; (2) the change in emphasis from that of correlation coefficients to that of standard error as the most useful measure of test reliability; (3) the inclusion of a section on standards for test use in the 1974 revision; (4) the conflict between test-takers' rights and test security; and, (5) the relationship between professional standards and public practice. (RL)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

The Huron Institute

NATIONAL CONSORTIUM ON TESTING PRO. IFC

123 MOUNT AUBURN STREET
CAMBRIDGE, MASSACHUSETTS 02138
617-661-9285

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

NATIONAL CONSORTIUM ON TESTING

STAFF CIRCULAR NO. 3

STANDARDS FOR TESTS AND TEST USE*

October 1978

Walt Haney
The Huron Institute
123 Mount Auburn Street
Cambridge, MA 02138

"PERMISSION TO REPRODUCE THIS
MATERIAL
HAS BEEN GRANTED BY

P. Howts

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

This is the third of a series of staff circulars which will be distributed as part of the Huron Institute's staff work for the National Consortium on Testing. Circulars will be distributed periodically over the next two years. This circular is intended to spur discussion on the topic of test standards and to raise important issues worthy of additional attention, not to resolve them. It is produced primarily as a resource document for member organizations of the NCT. Although the circular has been prepared as part of staff work for the Consortium, it does not necessarily reflect the views of the Consortium nor of individual Consortium members.

* The author would like to thank David K. Cohen, David Wiley, Elinor Woods, Judah Schwartz, and Mel Novick for helpful comments on a draft version of this paper. The paper does not, of course, necessarily reflect their views nor the views of anyone other than the author.

CONTENTS

	Page
I. Introduction	1
II. Ethical Standards	3
• 1953 Ethical Standards	4
• 1963 Ethical Standards	6
• 1977 Ethical Standards	6
III. Technical Standards	10
• 1954 Recommendations	10
• 1966 Standards	13
• 1974 Standards	17
IV. Conclusions	28
• Validity	30
• Reliability	30
• Test Use	32
• Test-takers' Rights and Test Security	33
• Professional Standards and Public Practice	35

I. Introduction

A high school newspaper carried a page-one headline: "Meet the geniuses of the incoming class" and then listed all pupils of IQ 120 and up with numerical scores. Then under a heading: "These are not geniuses, but good enough" were listed all the rest, with IQ's down to the 60's.

* * * *

A new battery of tests for reading readiness was introduced in a school. Instead of the customary two or three, 12 beginners, were this year described by the test as not ready for reading. They were placed in a special group and given no reading instruction. The principal insisted that if the parents or any one else tried to teach them to read "Their little minds would crack under the strain." In at least two cases parents did teach them to read with normal progress in the first semester, and later mental tests showed IQ's above 120. (APA, 1953)

Efforts to develop standards for tests and testing have a long history. As far back as 1895, a committee of the American Psychological Association (APA) was appointed to investigate the feasibility of standardizing the collection of mental and physical measurements. After the turn of the century, several different APA committees were set up with the goal of standardizing psychological tests, and evaluating the relative merits of different kinds of assessment devices (Singleton, n.d.). Individual psychologists did prescribe specific standards for tests--for example in 1924 Truman Kelley wrote that a test needed a reliability of 0.94 to be useful in evaluating levels of individual accomplishment (Kelley, 1924)--but organized efforts to develop professional standards for tests bore little fruit until mid-century.

By that time, the growing popularity of psychological testing and incidents of test abuse such as those quoted above apparently prompted

the APA to adopt a set of ethical principles concerning testing. Around the same time, the APA and other groups also developed technical recommendations regarding tests (APA, AERA, NCME, 1954). Since then, the APA ethical principles have been regularly revised and the technical standards periodically updated.*

The purpose of this paper is to identify the broad changes in these sorts of standards. The premise behind the paper is that the history of past efforts to develop testing standards can inform future attempts to promote reasonable use of tests. In particular, it is hoped that this paper will stimulate thinking of National Consortium on Testing members regarding the possibility of developing and promoting the use of test standards.

The paper is organized into four sections. After this introduction, the second section traces the way tests and testing were treated in three major versions of the APA's Ethical Standards of Psychologists. The third section recounts the evolution of technical standards through three major versions. The fourth section summarizes the manner in which these two sets of standards--ethical and technical--evolved and offers several suggestions about the issues which the two sets of standards do and do not treat.

* There is a third set of professional standards which also deal with testing, namely the Ethical Standards of the American Personnel and Guidance Association (APGA). One of six sections in these standards deals specifically with testing. However, since these standards are of more recent vintage and give deference to some of the other testing standards discussed here, the APGA standards will not be included in this analysis. For more information see APGA, 1961, 1965, 1972, 1976.

Before launching into the text of this paper, let us pause briefly to consider what is meant by the word "standard." Apart from specialized usages, as in botany, economics and music, the word has four general meanings which are pertinent to this discussion. Standard can mean a level of excellence or attainment regarded as a measure of adequacy. Standard can also refer to something established as a rule or basis of comparison in measuring quality or value; it may also mean a model or example commonly accepted or adhered to, that is a criterion set for usage or practice, as in moral standards. And in an older meaning, a standard may be a figure or object used as an emblem or symbol of a person or a group of people. Which of these meanings is implied in the term "test standards" is not altogether clear. Professional standards for tests and test use seem, as we shall see, to encompass all of these different meanings; meanings with very different implications for how such standards might be viewed.

II. Ethical Standards of Psychologists.

Arlene Kaplan-Daniels (1973) has observed that psychologists "have expended considerable effort to describe and give advice about ethical problems to their members..." (pp. 45-46). The first formal code of ethics for psychologists was adopted by the APA in 1952. The latest version of ethical standards for psychologists was adopted by the APA Council of Representatives in January 1977. In the intervening twenty-five years the ethical standards evolved through several versions and numerous amendments. Without tracing the details of this evolution, this section will compare the three major versions of the ethical standards: the original 1953 standards (APA 1953); the 1963 standards (APA 1963, 1967); and the most recent 1977 set of standards (APA 1977).

The analysis is limited mainly to those portions of the standards dealing directly with psychological testing.

1953 Ethical Standards. The original 1953 code of ethics of psychologists was presented in two volumes: one was a detailed statement "for use primarily by professional persons," and the other was a summary of the salient principles of the larger document. The detailed document was remarkably lengthy--around 170 pages, divided into six sections, each containing statements of general problem areas, examples of specific incidents involving ethical behavior, and specific principles relevant to each area.

One subsection of the 1953 Standards specifically concerned "publishing and using psychological tests and diagnostic aids." This subsection was divided into seven problem areas all relating to "the sale and distribution of tests and diagnostic aids." No attempt was made "to cover the ethics or standards of test construction, or test application, although the distinctions are at times difficult to make." (p. 143). The seven problem areas concerned:

- 1) the qualifications of tests users (3 principles);
- 2) the role of a psychologist sponsoring test activities (4);
- 3) the role and qualifications of test publishers' representatives (3);
- 4) choice of distributor for a new test (1);
- 5) readiness of a test for release (1);
- 6) the description of tests in manuals and publications (5); and
- 7) safeguarding testing materials (2).

Altogether, nineteen (19) out of more than 100 principles dealt specifically with testing. The first of these nineteen principles, regarding the qualifications of test users, stated:

Professional standards require that tests and diagnostic aids be released only to persons who can demonstrate that they have the knowledge and skill necessary for their effective use and interpretation. Tests can be classified in the following categories:

Level A Tests or aids which can be adequately administered, scored and interpreted with the aid of the manual and a general orientation to the kind of organization in which one is working. Examples: educational achievement, trade, and vocational proficiency tests.

Level B Tests or aids which require some technical knowledge of test construction and use, and of supporting psychological and educational subjects such as statistics, individual differences, the psychology of adjustment, personnel psychology, and guidance. Examples: general intelligence and special aptitude tests, interest and personality inventories....

Level C Tests and aids which require substantial understanding of testing and supporting psychological subjects, together with supervised experience in the use of these devices. Examples: clinical tests of intelligence, personality tests, and projective tests. (Principle 541-1; APA, 1953, pp. 146-7.)

Other principles regarding testing elaborated on the need for limiting the distribution of psychological tests. "Good professional practice demands that test materials be retained by students only if they are graduate students in fields in which tests are professional equipment,..." (Principle 5.42-2, p. 149). "Instructors are expected to manage the use of psychological tests and other devices, the use of which might be spoiled by familiarizing the general public with their specific contents or underlying principles, in such a way as to limit access to them to persons who have a professional interest and who will safeguard their use" (Principle 5.47-1, p. 156). In short, the 1953 Ethical Standards gave considerable attention to the need for safeguarding psychological tests as "professional equipment," both to prevent misuse and to avoid their invalidation.

1963 Ethical Standards. By 1963, the ethical principles had shrunk in size. According to a later publication, "On the basis of accumulated experience, it became possible to distill from the original [1953] code, a set of more general principles" (APA, 1967, p. vii). The 1963 standards, including amendments, consisted of nineteen relatively general principles. Only three pertained directly to testing; specifically to test security, interpretation and publication:

Test Security. Psychological tests and other assessment devices, the value of which depends in part on the naivete of the subject, are not reproduced or described in popular publications in ways that might invalidate the techniques. Access to such tests is limited to persons with professional interests who will safeguard their use.

Test Interpretation. Test scores, like test materials, are released only to persons who are qualified to interpret and use them properly.

Test Publication. Psychological tests are offered for commercial publication only to publishers who present and represent their tests in a professional way and distribute them only to qualified users. (APA, 1967, pp. 51, 54, 56)

All three of these ethical principles pertaining to testing emphasized the need to control distribution of test information. The reasons for such control seemed much the same as in the previous decade. Such restrictions of test information to qualified professional users was necessary, went the apparent rationale, both to ensure proper use and to prevent invalidation of instruments due to uncontrolled dissemination.

1977 Ethical Standards. By 1977, the ethical principles had shrunk still further in number. Overall, the nineteen principles in the 1963 Ethical Standards were reduced to just nine principles in 1977. The three 1963 principles relating specifically to testing were merged in

1977 into a single principle concerning utilization of assessment techniques:

In the development, publication, and utilization of psychological assessment techniques, psychologists observe relevant APA standards. Persons examined have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgments were based. Test users avoid imparting unnecessary information which would compromise test security, but they provide requested information that explains the basis for decisions that may adversely affect that person or that person's dependents.

a. The client has the right to have and the psychologist has the responsibility to provide explanations of the nature and purposes of the test and the test results in language that the client can understand, unless, as in some employment or school settings, there is an explicit exception to this right agreed upon in advance. When the explanations are to be provided by others, the psychologist establishes procedures for providing adequate explanations.

b. When a test is published or otherwise made available for operational use, it is accompanied by a manual (or other published or readily available information) that fully describes the development of the test, the rationale, and evidence of validity and reliability. The test manual explicitly states the purposes and applications for which the test is recommended and identifies special qualifications required to administer the test and to interpret it properly. Test manuals provide complete information regarding the characteristics of the normative population.

c. In reporting test results, psychologists indicate any reservations regarding validity or reliability resulting from testing circumstances or inappropriateness of the test norms for the person tested. Psychologists strive to insure that the test results and their interpretations are not misused by others.

d. Psychologists accept responsibility for removing from clients' files test score information that has become obsolete, lest such information be misused or misconstrued to the disadvantage of the person tested.

e. Psychologists offering test scoring and interpretation services are able to demonstrate that the validity of the programs and procedures used in arriving at interpretations are based on appropriate evidence. The public offering of an automated test interpretation service is considered as a professional-to-professional consultation. The psychologist makes every effort to avoid misuse of test reports. (APA, 1977, p. 6).

In sum, the Ethical Standards of Psychologists seem to adhere most closely to the meaning of standard as a model or example commonly accepted or adhered to as a criterion for acceptable practice. Indeed, the 1967 publication of a Casebook on Ethical Standards of Psychologists suggests such an interpretation. The Casebook contains selected cases from 1959 to 1962 "illustrative of the principles involved in ethical problems..." (APA, 1967, p. ix). In this respect, the materials in the Casebook are quite similar to the ethical incidents included in the original 1953 Ethical Standards. In addition, however, the 1967 Casebook also describes the application of ethical standards to these actual cases by the APA's Committee on Scientific and Professional Ethics and Conduct. In several instances, the conduct of individuals involved in these cases was judged to be unethical, and as a result they were dropped from membership in the APA. So the Ethical Standards have indeed served in at least some instances as criteria for acceptable practice.

Nevertheless, there have been substantial changes in the ethical standards for psychologists over the last quarter-century. First, the sheer number of stated principles has dropped sharply. The original 1953 version of the standards listed more than 100 ethical principles. In the 1963 version, these were distilled to yield a set of nineteen more general principles. The distillation went further in the 1977 version which contained just nine general principles to guide the ethical conduct of psychologists. The distillation process was evident in the number of ethical principles directly relevant to testing. The 1953 Ethical Standards listed nineteen principles regarding testing. In the 1963 Standards, these were reduced to three principles, relating to test

security, interpretation and publication; the 1977 version contained a single general principle concerning utilization of assessment techniques. In large part the diminution in numbers of principles represented a distillation of specific principles into more general ones. For example, the 1953 document prescribed specific A, B & C levels of tests which should be released only to people with specific sorts of training. The later versions of the Ethical Standards simply stated without elaboration that tests should be released only to persons qualified to interpret them properly. In addition to this distillation of specific guidelines into fewer more general principles, another more qualitative change is apparent in the latest version of the Ethical Standards. Specifically, the 1977 Standards give much more attention to the rights of psychologists' clients, and specifically to the rights of persons tested, than did earlier versions of the Ethical Standards. Earlier versions gave stress mainly to the responsibilities of psychologists as scientists and professional persons. To this perspective, the 1977 Standards added emphasis to the rights of individuals served by psychologists. With respect to psychological testing, this attention to consumers' rights led to an apparent contradiction. Concerning testing, the 1977 Standards states that "Persons examined have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgements were made." In light of the very next sentence, the modifier "where appropriate" looms large and uncertain: "Test users avoid imparting unnecessary information which would compromise test security...." An obvious question remains: When do the rights of test takers leave off and the need for test security begin?

III. Technical Standards

At the end of the 1977 Ethical Standards, it is stated that "psychologists are responsible for knowing about and acting in accord with the standards and positions of the APA," as represented in such official documents as the 1974 version of the Standards for Educational and Psychological Tests.^{*} The 1974 version of these standards is an adaptation of technical standards on testing dating back to the early 1950s. The paragraphs below recount how the technical standards evolved through three major versions.

1954 Recommendations. In the early 1950s the APA, the American Educational Research Association (AERA) and the National Council on Measurements Used in Education (the forerunner of NCME) collaborated to produce in 1954 a set of Technical Recommendations for Psychological

^{*} The 1977 Ethical Standards directed adherence also to eight other documents:

- American Association of University Professors. Statement on Principles on Academic Freedom and Tenure. Policy Documents and Report, 1977, 1-4.
- American Psychological Association. Guidelines for Psychologists for the Use of Drugs in Research. Washington, D. C.: Author, 1971.
- American Psychological Association. Principles for the Care and Use of Animals. Washington, D. C.: Author, 1971.
- American Psychological Association. Guidelines for conditions of employment of psychologists. American Psychologist, 1972, 27, 331-334.
- American Psychological Association. Guidelines for psychologists conducting growth groups. American Psychologist, 1973, 28, 933.
- American Psychological Association. Ethical Principles in the Conduct of Research with Human Participants. Washington, D. C.: Author, 1973.
- American Psychological Association. Standards for Providers of Psychological Services. Washington, D. C.: Author 1977.
- Committee on Scientific and Professional Ethics and Conduct. Guidelines for telephone directory listings. American Psychologist, 1969, 24, 70-71

Tests and Diagnostic Techniques.** The 1954 Technical Recommendations were based on the "essential principle" that "a test manual should carry information sufficient to enable any qualified user to make sound judgements regarding the usefulness and interpretations of the test" (p. 2). Altogether, the document contained more than 160 separate standards grouped into six categories, pertaining to:

- A. Dissemination of Information (10 standards)
- B. Interpretation (18)
- C. Validity (66)
- D. Reliability (31)
- E. Administration and Scoring (10)
- F. Scales and Norms (29)

Unlike the recent versions of the Ethical Standards, the technical recommendations were differentiated in two ways. First, they were differentiated into major and subsidiary recommendations (with major ones printed in boldface type and labelled with whole numbers, for example A3; and subsidiary standards printed in regular-face type and numbered in decimal form, for example A3.1 and A3.11). Second, each recommendation was rated as essential, very desirable, or desirable. The explanation for these three levels of recommendations went as follows:

Manuals can never give all the information that might be desirable, because of economic limitations. At the same time, restricting this statement of recommendations to essential information might tend to discourage reporting of additional information. To avoid this, recommendations are grouped in three levels: ESSENTIAL, VERY DESIRABLE, and DESIRABLE....

The statements listed as ESSENTIAL are intended to be the consensus of present-day thinking as to what is normally required for operational use of a test....

** For a more complete account of the development of the 1954 Technical Recommendations, see Singleton, (n.d.).

The category VERY DESIRABLE is used to draw attention to types of information which contribute greatly to the user's understanding of the test....

The category DESIRABLE includes information which would be helpful.... (pp. 5-6).

With these distinctions, the authors of the Technical Recommendations seemed to be attempting to bridge two meanings of the term "standard." Considering only essential recommendations, we would seem to have a statement of attainment in test development that could be regarded as a measure of adequacy. However, if we took the recommendations more broadly including very desirable and desirable as well as essential ones, we would seem to have a broader basis of comparison for measuring the quality of test development more generally.

Nevertheless, the authors of the 1953 Recommendations seemed to have fairly flexible ideas regarding what it was they had produced. At various points it was suggested that the 1953 document was:

- a statement representing a consensus as to what information is most helpful to the test consumer (p. 1),
- a guide to test development and reporting (p. 6),
- a checklist of factors to consider in designing standardization and validation studies (p. 6),
- a reminder regarding features to be considered in choosing tests for a particular program (p. 7), and
- a statement to be used by individual members of the professions to improve their own work (p. 7).

Despite these varied purposes, the authors also pointed out some things that the document was not. The Recommendations did not attempt to set out "minimal statistical specifications" for tests (p. 2). Authors of the document were careful to point out too that "the present recommendations apply

to devices which are distributed for use as a basis for practical judgements rather than solely for research" (p. 4). The authors also opined:

Almost any test can be useful for some functions and in some situations. But even the best test can have damaging consequences if used inappropriately. Therefore, ultimate responsibility for improvement of testing rests on the shoulders of test users. (p. 7).

Given this sense of the locus of responsibility for testing practices, it was not surprising that the authors of the Technical Recommendations intended them to be used "without reference to any enforcement machinery" (p. 7).

1966 Standards. In 1966, the APA, AERA and NCME issued a revised version of the technical recommendations called Standards for Educational and Psychological Tests. In format and in length the 1966 Standards were very similar to the 1954 Recommendations. The 1966 version contained 160 separate standards, differentiated into major and subsidiary ones, with each rated as essential, very desirable, or desirable, and grouped into the same six sections as the 1954 standards:

- A. Dissemination of Information (12 standards)
- B. Interpretation (18)
- C. Validity (60)
- D. Reliability (34)
- E. Administration and Scoring (14)
- F. Scales and Norms (22)

The degree of change between corresponding sections of the 1954 Recommendations and the 1966 Standards varied markedly, however. The first two sections remained essentially the same, but the sections on

validity and reliability were revised considerably. The section of the 1954 Technical Recommendations regarding validity had begun:

Validity information indicates to the test user the degree to which the test is capable of achieving certain aims. Tests are used for several types of judgement, and for each type of judgement a somewhat different type of validation is involved. We may distinguish four aims of testing:

1. The test user wishes to determine how an individual would perform at present in a given universe of situations of which the test situations constitutes a sample.
2. The test user wishes to predict an individual's future performance (on the test or on some external variable).
3. The test user wishes to estimate an individual's present status on some variable external to the test.
4. The test user wishes to infer the degree to which the individual possesses some trait or quality (construct) presumed to be reflected in the test performance (p. 13).

Corresponding to these four types of uses, the 1954 Technical Recommendations defined four types of validity: content validity, predictive validity, concurrent validity and construct validity.

In the 1966 section on validity, the Standards listed just three aims of testing (p. 12), and correspondingly, only three types of validity. Predictive validity and concurrent validity were treated in 1966 as alternative forms of criterion-related validity.

The apparent reason for the change, that is the lumping together of predictive and concurrent validities as different forms of criterion-related validity, was that the two are conceptually the same. The only difference is the time at which the criterion data are gathered. If gathered at roughly the same time as the test, the criterion data would provide an estimate of concurrent validity, but if gathered at a

later point in time, they would provide information on predictive validity.

Changes between 1954 and 1966 concerning test reliability were somewhat sharper. The 1954 document had described reliability almost entirely in terms of correlation coefficients.

The coefficient of internal consistency [refers] to a measure based on internal analysis of data obtained on a single trial of a test....A correlation between scores from two forms given at essentially the same time we shall refer to as a coefficient of equivalence. The correlation between test and retest, with an intervening period of time, is a coefficient of stability. Such a coefficient is also obtained when two forms of the test are given with an intervening period of time (p. 28).

The 1966 Standards continued to discuss reliability in terms of comparability of forms, internal consistency and comparisons over time. In a sharp change, however, the 1966 document reported that the classification of "coefficients into several types (e.g., coefficient of equivalence) has been discarded. Such a terminological system breaks down as more adequate statistical analyses are applied and methods are more adequately described" (p. 26). Instead of defining reliability strictly in terms of correlation coefficients, the 1966 standards recommended that "the estimation of clearly labeled components of error variance is the most informative outcome of a reliability study" (p. 26, emphasis in original). The derogation of the correlation coefficient as a measure of reliability reflected a rapprochement between the Pearsonian correlational approach to test theory and the Fisherian experimental design and analysis tradition of statistics (Stanley, 1971, pp. 371-2). The basic argument against the correlation coefficients as the prime measure of reliability is that in terms of classical test theory it allocates

all sources of observed score variance to true score (the numerator of the fraction from which the coefficient is calculated) or error score (the denominator minus the numerator). The problem is that there are far more than two sources of variation for test scores. A specific example of this general problem is that correlation coefficients of reliability can be dramatically affected by the range of scores on which they are calculated. The correlation coefficient was apparently de-emphasized as a measure of reliability, and the components of variance approach emphasized, precisely because of such problems.

Compared to the treatment of validity and reliability issues, other sections of the 1966 Standards changed relatively little. The sections on "administration and scoring" and on "scales and norms" differed relatively little from the corresponding sections of the 1954 standards. Several principles were reworded and some were reorganized but the gist of these two sections remained essentially the same.

The "essential principle" behind the 1966 standards remained the same as the one stated in the 1954 version. And the latter version repeated the disclaimer concerning appropriateness of setting minimum statistical specifications for tests (p. 2). Also, the 1966 document repeated a warning from the 1954 Recommendations that a test manual "is to be judged not merely by its literal truthfulness, but by the impression it leaves on the reader" (p. 2).

Along these lines, however, the 1966 standards evidenced little effort to bridge a curious gap in the 1954 document. In the 1954 version, it was asserted as noted above, that the "ultimate responsibility for improvement of testing rests on the shoulders of test users" (p. 7). Yet the Technical Recommendations were "intended to guide test development

and reporting" (p. 6). Scarcely a word of advice was offered directly to test users--on whose shoulders was placed the "ultimate responsibility" for improvement of testing. In the main, this discrepancy remained in the 1966 Standards. The closing lines of the introduction to the 1966 Standards reiterated a caveat from the 1954 document:

It is important to note that publication of superior information about tests by no means guarantees that tests will be used well. The continual improvement of courses which prepare test users and of leadership in all institutions using tests is a responsibility in which everyone must share (p. 6).

1974 Standards. By the early 1970s, "it had become apparent that problems and issues in testing, especially problems in selection for employment or for admission to educational institutions, had outdated the 1966 publication" (APA, AERA, NCME, 1974, p. 2). Consequently, a joint committee of the APA, the AERA and NCME was convened to revise the standards.

After three years of work, and numerous committee meetings, the revised standards were formally approved by the three organizations and were published in 1974. The new standards were both a revision and an extension of the earlier version. In particular, the 1974 document attempted to bridge the gap between test manuals and test use. It was intended "to guide both test developers and test users" (p. 1).

Consequently, the 1974 technical standards had a substantially different format than earlier versions. The rating of standards as essential, very desirable or desirable was maintained, but the domain of the 1974 standards was expanded considerably and organized into three major parts:

Standards for Test Manuals and Reports

- A. Dissemination of Information (16 standards)
- B. Aids to Interpretation (20)
- C. Directions for Administration and Scoring (14)
- D. Norms and Scales (28)

Standards for Reports of Research on Reliability and Validity

- E. Validity (82)
- F. Reliability and Measurement Error (26)

Standards for the Use of Tests

- G. Qualifications and Concerns of Users (9)
- H. Choice or Development of Test or Method (9)
- I. Administration and Scoring (16)
- J. Interpretation of Scores (24)

Altogether the 1974 standards were expanded by more than 50%, from 160 specific standards in 1966 to more than 240 in 1974. Most of the new standards were added in the third part, concerning use of tests, but there were also substantial revisions in the topics covered in earlier versions of the standards.

In section A a new group of standards was added calling for inclusion in the test manual of a description of the development of the test. New standards were added to section B calling for the test manual to "describe clearly the psychological, educational or other reasoning underlying the test and the nature of the characteristic it is intended to measure" (B3, p. 15), and prescribing that "computer services for test interpretation should provide a manual reporting the rationale and evidence in support of computer-based interpretations of scores" (B6, p. 7). Dropped from the standard on qualifications for test administration and interpretation was any reference to level A, B & C tests--the distinction drawn originally in the 1953 ethical standards, and referred to in both the 1954 Recommendations and the 1966 Standards.

Section C on directions for administration and scoring was mainly the same as the old section E in the 1966 standards. One new standard (C2), concerning preparation of the examinee for the examination was

added to this section, but it received the lowest rating of desirable. In section D on norms and scales, most standards were revised somewhat and reordered and a new standard concerning group norms (D7) was added.

The section on validity was revised considerably over the 1966 standards. The section still discussed content, construct and criterion-related validities, but unlike the 1966 Standards, criterion-related validity was given far more prominence than the other two forms of validity. Indeed, in the 1974 standards, more than 60% of the standards in the validity section dealt with criterion-related validity. It seems likely that this emphasis was due mainly to the concern over using tests in personnel selection that motivated the 1974 revision. Predictive validity of a selection test in terms of job performance was at that time one of the major criteria by which the legal propriety of employment selection was judged (Lerner, 1978).

In expanding the treatment of criterion-related validity, several new standards were added. Doubtless the most significant was a new essential standard concerning test bias:

A test user should investigate the possibility of bias in tests or in test items. Wherever possible, there should be an investigation of possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. (E9, p. 43).

The addition of this standard concerning investigation of test bias was significant but its placement in the section on criterion-related validity was peculiar. Concern over test bias would seem equally applicable to construct validity--especially in light of the long history

of controversy over bias in tests of the construct of intelligence. It seems likely that the placement of the bias standard is explainable only in terms of the apparent preoccupation of the authors of the 1974 Standards with criterion-related validity.

In contrast to the treatment of validity issues, the section on reliability and measurement error was somewhat smaller in the 1974 Standards than the corresponding section of the 1966 Standards. The structure of the new version on reliability was, however, much the same as that of the earlier version. There was some shrinkage in the detailed standards and some rewording of major standards, but it is easy to identify an exact correspondence between the six major (i.e. printed in bold-faced type) reliability standards in the 1974 standards and the 1966 reliability standards. Nevertheless, there does seem to be an important shift in emphasis in the wording of the 1974 standards on reliability. Specifically, this latest version of the standards, gives standard error as an index of reliability more prominence relative to reliability coefficients than did either the 1954 or the 1966 standards. This shift in emphasis seems quite important so I will return to this point in the concluding section of this paper.

The third part of the 1974 standards, "standards for the use of tests" was completely new. It was divided as previously mentioned, into four sections: 1) qualification and concerns of users; 2) choice or development of test or method; 3) administration and scoring; and 4) interpretation of scores. The major principles in these four categories are summarized in Table 1.

Regarding qualifications of test users the 1974 Standards took a very different tack than had been taken in the 1953 Ethical Standards. No mention at all was made of level A, B & C tests. Instead, the 1974 principles took the more vague approach of asserting that test users, whoever they might be, should have training or guidance appropriate to the tests they use, without specifying what training would be appropriate to different types of tests.

Regarding choice of development of test or method, the 1974 standards made the sensible suggestion that "test users should not use habitually the same test or method of assessment for all purposes; neither should they assess only those characteristics that are easily or conveniently assessed and fail to consider other, possibly more important, characteristics" (p. 60).

The standards regarding administration and scoring for test users seem quite straightforward. They are premised on the idea that "the basic principle is standardization, when decisions are based on test scores, the decision for each individual should be based on data obtained under circumstances that are essentially alike for all" (p. 64). The major problem with this premise is that it assumes that circumstances that are objectively alike in all essential respects will yield equally meaningful test results. Clearly this will not necessarily be the case. For example, if a test is administered to a person with impaired vision in the same manner as for a person with 20-20 eyesight, the results will not necessarily be equally meaningful. The other aspect of the administration and scoring standards seeming to get unwarrantedly small attention concerns test security. We shall return to this issue in the final section of this paper.

Table I: 1974 STANDARDS FOR THE USE OF TESTS*

G. Qualifications and Concerns of Users

- G1. A test user should have a general knowledge of measurement principles and of the limitations of test interpretations. **Essential.**
- G2. A test user should know and understand the literature relevant to the tests he uses and the testing problems with which he deals. **Very Desirable.**
- G3. One who has the responsibility for decisions about individuals or policies that are based on test results should have an understanding of psychological or educational measurement and of validation and other test research. **Essential.**
- G4. Test users should seek to avoid bias in test selection, administration, and interpretation; they should try to avoid even the appearance of discriminatory practice. **Essential.**
- G5. Institutional test users should establish procedures for periodic internal review of test use. **Essential.**

H. Choice or Development of Test or Method

- H1. The choice of development of tests, test batteries, or other assessment procedures should be based on clearly formulated goals and hypotheses. **Essential.**
- H2. A test user should consider more than one variable for assessment and the assessment of any given variable by more than one method. **Essential.**
- H3. In choosing an existing test, a test user should relate its history of research and development to his intended use of the instrument. **Essential.**
- H4. In general a test user should try to choose or to develop an assessment technique in which "tester-effect" is minimized, or in which reliability of assessment across testers can be assured. **Essential.**
- H5. Test scores used for selection or other administrative decisions about an individual may not be useful for individual or program evaluation and vice versa. **Desirable.**

I. Administration and Scoring

- I1. A test user is expected to follow carefully the standardized procedures described in the manual for administering a test. **Essential.**
- I2. The test administrator is responsible for establishing conditions, consistent with the principle of standardization, that enable each examinee to do his best. **Essential.**

* This table lists only major standards for test use, that is standards printed in bold-faced type. In the full version of the 1974 Standards, most of these major standards are followed by comments and subsidiary standards.

- I3. A test user is responsible for accuracy in scoring, checking, coding, or recording test results. Essential.
- I4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. Essential.
- I5. The test user shares with the test developer or distributor a responsibility for maintaining test security. Essential.

J. Interpretation of Scores

- J1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. Essential.
- J2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the recipient to interpret them correctly. Desirable.
- J3. The test user should recognize that estimates of reliability do not indicate criterion-related validity. Essential.
- J4. A test user should examine carefully the rationale and validity of computer-based interpretations of test scores. Essential.
- J5. In norm-referenced interpretations, a test user should interpret an obtained score with reference to set of norms appropriate for the individual tested and for the intended use. Essential.
- J6. Any content-referenced interpretation should clearly indicate the domain to which one can generalize. Essential.
- J7. The test user should consider alternative interpretations of a given scores. Essential.
- J8. The test user should be able to interpret test performance relative to other measures. Very Desirable.
- J9. A test user should develop procedures for systematically eliminating from data files test-score information that has, because of the lapse of time, become obsolete. Essential.

The major principles concerning users' interpretations of test scores were summarized in Table 1. One of them seems especially important. Essential principle J1 states:

A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. (p. 68)

On its surface this principle seems to preclude the interpretation of test scores as indicators of innate intelligence or fixed ability. Since we have no casebook of interpretations of the 1974 Standards such as is available for the 1963 Ethical Standards, the only aid available in interpreting the meaning of the above principle is the subsidiary principle and comments appended to it:

J1.1. A test user should consider the total context of testing in interpreting an obtained score before making any decisions (including the decision to accept the score). Essential.

[Comment: The standard is that one must avoid the abdication of responsibility by relying exclusively on an obtained score. Users should, in particular, look for contaminating or irrelevant variables that may have influenced obtained scores; for example, in testing to classify school children, scores may be influenced by behavior problems, visual or hearing defects, language problems, and racial or cultural factors, as well as by ability.] (p. 68)

These elaborations seem to uphold the interpretation suggested above. Principle J1 of the 1974 Standards apparently proscribes the interpretation of a test score as a measure of fixed intelligence or ability unless specific consideration is given at least to the variables mentioned above "that may have influenced obtained test scores."

The new section of the 1974 Standards concerning use of tests

seems to be a clear step forward from the 1954 and 1966 standards. For the first time the standards made a clear attempt to provide guidance to those upon whose shoulders had been placed the "primary responsibility for the improvement of testing." (p. 7). Nevertheless, there is a major problem in the new section on test use. The implications of alternative uses of tests are not very clearly spelled out, and the main distinction drawn regarding differential application of standards on use seems somewhat out-of-touch with the realities of how tests are actually used to contribute to decision-making.

The introduction to the part of the 1974 Standards on "Standards for Test Use," begins "There are many kinds of test use. As one example, test scores are used for decisions to select or to reject applicants for jobs, schools, or other opportunities," (p. 56). The second paragraph notes that "Another use is as a prescriptive aid where different scores imply different treatments." The third paragraph relates that "A test score may be used to certify that an individual has met some designated standard, that a person is qualified to perform certain skilled tasks, that a child is qualified for a remedial program, or that a defendant can stand trial."

Although these alternative uses of tests--in selection, prescription and certification--are briefly discussed, the 1974 Standards do not clearly distinguish how or even if different standards might apply to these alternative uses. In only one rather vague respect does the 1974 document suggest that different uses of tests require differential application of standards for use. The Standards suggest that tests may contribute to making decisions about individuals or to making institutional decisions about programs, funding or other organizational issues.

As examples of using tests at the individual levels, the 1974 Standards mentions use of test scores by a "counselor to help a student make a vocational choice or to help a couple in marriage counseling communicate more clearly with each other" (p. 56). Concerning such use, the 1974 Standards remarks:

These are clinical, diagnostic, and individualistic uses of tests in a continuing relationship between a test user and an individual. Because the relationship is a continuing one, tentative decisions or judgments can be modified as new information is accumulated. (p. 57).

Regarding use of tests at the aggregate level, the Standards comment:

Test scores may constitute the dependent variable or criterion measure in an institutional research study. A program may be continued or terminated on the basis of test results; an institution may receive more or less funding because of test results; test results may be considered in organizational analysis or in making program changes. The Standards do not deal fully with these problems; their emphasis is more on the interpretation of scores of individuals. A companion volume is planned dealing with standards for test use in program evaluation, policy-related research, and curriculum evaluation; it will also address issues of research design and of data analysis (pp. 56-57).

As a result of this concern, "the suggestion for a separate publication on guidelines for program evaluation was circulated to a number of persons having expertise in this area. It was decided that there was, indeed a need for such a document and that an advisory group should be appointed to carry out this task" (Singleton, n.d., p. 11). A committee to draft program evaluation guidelines was appointed and is chaired by Daniel Stufflebean of Western Michigan University. A draft set of program evaluation guidelines has been prepared by this committee, but has not yet been made available for unrestricted public review.

The 1974 document maintained that "The standards in the present volume are to varying degrees directed to all forms of use" (p. 57). Nevertheless, it was suggested that the standards might be differentially applied at the individual and aggregate levels:

As the use of tests moves along a continuum from the description of a single individual, in a situation allowing for corrections of erroneous interpretations, to making decisions about large numbers of people, the test user must apply more of the standards and perhaps, apply them more rigorously (p. 57).

This recommendation regarding differential application of standards for test use seems rather odd. First it seems to ignore the reality that decisions at the aggregate level, for example, regarding programs or funding are rarely, if ever, based strictly on test results (see Cohen and Garet, 1975, for example). Decisions about programs and funding usually take place in the political arena, "in a situation allowing for corrections of erroneous interpretations" of test results. Decisions about individuals on the basis of test scores, on the other hand, even when drawn in clinical settings do not necessarily entail mechanisms for correcting erroneous interpretations. Recall, for example, the two incidents quoted at the start of this paper. Also the literature on special education suggests that individual children often do not encounter the situation described in the 1974 Standards, namely ones "allowing for corrections of erroneous interpretations." In a special volume on Issues in the Classification of Children, Goldstein et al (1975), for example, wrote:

In spite of all that has been written lately about the pitfalls of this system [of special education in the schools], the pattern of routine testing, labeling and placement of children persists; in the vast majority of cases,

the adaptive behavior of the children in settings other than the classroom is never evaluated....while the processing of labeling is a formal procedure, that of removal of labels is not....The child usually has no advocate. (pp. 33, 36, 37)

In sum it seems as though institutions are usually better equipped than individuals to protect themselves against misuse of test results. On this point simply recall the revealing assertion in the 1963 ethical standards that the value of tests "depend in part on the naivete of the subject." If this is so, it clearly would be appropriate to require that standards for test use with individuals be applied fully as rigorously as those with respect to groups--or maybe more so.

IV. Conclusions

This brief review of the evolution of ethical and technical standards for testing has identified several broad changes in standards over the past 25 years and has suggested specific issues that seem to need more attention. In conclusion I summarize some of the broader changes in the standards and briefly discuss five specific issues arising from this review: (1) validity; (2) reliability; (3) test use; (4) test-takers rights and test security; and (5) the relationship between professional standards and public practice.

The most apparent change in the two sets of standards over the last quarter-century is that while the ethical standards directly relevant to testing have diminished in number, the technical standards have multiplied. The 1953 Ethical Standards listed 19 principles specifically relevant to tests and diagnostic aids. The 1963 version listed only three, and the 1977 Ethical Standards listed one. In contrast, the technical standards grew from around 160 in the 1954

Technical Recommendations and the 1966 Standards, to more than 240 in the 1974 Standards. Clearly there has been an inverse relationship between numbers of standards for ethics and of those for techniques. One might interpret this finding as evidence of the technocratic tendencies of professionals. After all, the very identity of professionals is probably based more on their possession of special knowledge or skills than on their possession of special ethics--though both are important in the ethos of professionalism. Also, however, the inverse relationship likely stems from a real difference between techniques and ethics. Techniques by their very nature are more precise and amenable to specific description, while ethics are more likely to be situation-bound and so best stated in more general terms. Authors of the later Ethical Standards may also, of course, have felt free to skip over certain specific testing issues because of the availability of the technical standards.

The second pattern evident in the comparison of these two sets of standards is that the rate of change in the technical standards seems to be accelerating faster than in the ethical standards. Major versions of the ethical standards appeared in 1953, 1963, and 1977. Major versions of the technical standards were issued in 1954, 1966, and 1974 and a special joint committee of the APA, AERA, and the NCME has recently recommended that the 1974 Standards should be revised. The accelerating rate of change of the technical standards has obvious implications for the standards' impact on testing practices. However, other specific issues also have clear relevance for practice. So before discussing this issue, let me consider them first.

Validity. In 1954 and 1966, the standards mentioned four types of test validity: content, predictive, concurrent and construct. However, in 1966 predictive and concurrent validity were grouped together as different forms of criterion-related validity and were given somewhat more emphasis than the other two forms of validity. This trend continued in the 1974 Standards with the bulk of space in the section on validity standards given to issues of criterion-related validity. The emphasis on criterion-related validity may have been due, as suggested above, to the concerns over use of tests in employee and educational selection that motivated the 1966 revision. Relatively little attention was given in the 1974 Standards to content validity (eight specific standards as opposed to 52 on criterion-related validity). The overwhelming emphasis on criterion-related validity seems particularly odd because the 1974 document states:

These [three] aspects of validity can be discussed independently, but only for convenience. They are inter-related operationally and logically; only rarely is one of them alone important in a particular situation. A thorough study of a test may often involve information about all types of validity (p. 26).

Reliability. Treatment of test reliability has evolved even more sharply through the three versions of the technical standards than has the treatment of validity. The 1954 Recommendations distinguished between reliability coefficients of internal consistency, equivalence of forms, and stability. The 1966 version discarded the classification of reliability coefficients into these types and instead recommended that "the

estimation of clearly labeled components of error variance is the most informative outcome of a reliability study" (p. 26). The 1974 Standards repeated this recommendation regarding the most informative outcome of a reliability study (p. 49), but also gave new prominence to the standard error as a measure of reliability.* The 1974 document advised:

Reliability coefficients have limited practical value for test users. The standard error of measurement is more useful; . . . Test users may use reliability coefficients in comparing tests, but they use standard errors of measurement in interpreting test scores. Information in a test manual about a standard error of measurement is ordinarily more important than information about a reliability coefficient (p. 50).

This change in the standards seems quite important. Standard errors are more useful than reliability coefficients not only because they may be more useful in interpreting test scores (for example in setting confidence intervals), but also because they help keep people's attention on what is being measured. Discussions of reliability in terms of correlation coefficients tend to become disembodied from the content of what tests measure, in part I suspect, because reliability coefficients are always unit-free numbers varying from 0.00 to 1.00. Standard errors, however, because they are, or at least can be, expressed in terms of the content of what is tested (for example in terms of numbers of items), help to root discussions in the content of the test at hand. In any event the upgrading of the standard error as a measure of test reliability seems to be a move in the right direction.

* The standard error can be derived from the relevant variance. The mathematical relationship between coefficients of reliability and variance components can be shown with models which decompose test scores in various components or facets of variance. See, for example, Cronbach et al. (1972).

Test Use. The inclusion of a section on standards for test use in the 1974 Standards also seems to be a step in the right direction. The 1966 Standards had proclaimed that the burden of responsibility for improving testing rests on the shoulders of test users, but had offered scant guidance concerning standards for use. The 1974 Standards moved to remedy this problem by including an entirely new section on test use. Several of the 1974 Standards on test use seem quite important. Standard J1 concerning users' interpretations of test scores seems especially so.

J1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances.

Both on its surface and in the commentary offered with respect to it, this standard seems clearly to proscribe the interpretation of test scores as measures of innate intelligence or fixed ability.

The main problem in the section of the 1974 Standards on test use is that it fails to elaborate on alternative uses of test results and to draw out the implications of alternative uses. In particular, the 1974 Standards seem to slight the need for rigorous adherence to standards on use when test results are used with individuals.

These problems may stem in part from the fact that the 1974 Standards attempts to serve so many ends. At various points, intentions are expressed that the 1974 Standards will:

- "describe in an explicit and conveniently available form the information required by test users" (p. 5);
- "state what type of studies should be completed before a test is ready for use" (p. 6);

- "provide useful guidelines for test users as well as test developers" (p. 6);
- "be used to extend the professional training of many test users who are not now being trained appropriately" (p. 7);
- "draw attention to recent developments in thinking about tests, test analysis, and test use" (p. 7);
- "provide a kind of checklist of factors to be considered in designing, standardizing, validating, scoring, and interpreting tests" (p. 7).

Is it even conceivable that a single 80-page document could serve all of these ends equally well?

Test-takers' rights and test security. Another aspect of test use that seems to be emerging as an increasingly important problem is the conflict between test-takers' rights and test security--or test secrecy as the issue is sometimes labeled outside the test standards. This issue appears more clearly in the 1977 Ethical Standards than in any of the technical standards. The nub of the problem is revealed in two sentences of the principle of the 1977 Ethical Standards on utilization of assessment techniques:

Persons examined have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgements were based. Test users avoid imparting unnecessary information which would compromise test security, but they provide requested information that explains the basis for decisions that may adversely affect that person or that person's dependents.

This statement reflects a consumers rights point of view far more strongly than any previous version of either the ethical or technical standards. Indeed, the degree of change on this point from the Ethical Standards in 1953 to those in 1977 probably reflects a great deal about how American society has changed in this quarter century.

Nevertheless, the phrase in the first sentence above, "where appropriate," still carries a mighty ambiguity. In discussing where it might be appropriate to give persons examined "the right to know the original data on which final judgements were based," in effect to give them the test, let us draw two distinctions. First, the case of test-takers' rights versus test security may be quite different in situations where people enter into the test-taking situation voluntarily with certain conditions specified a priori (for example, with aspirants to law schools taking the LSAT) and in those where individuals are in effect forced into taking a test (for example in a school's regular standardized testing program). Second, the case for test security prior to test administration seems fundamentally different than the case after a test has been given. The case for test secrecy prior to administration seems considerably stronger than the case afterwards. Calls for disclosure of standardized tests' contents after administration have, in fact, been embodied in recent legislative proposals regarding freedom of information concerning standardized testing in the U.S. Congress, in New York, and in California.* Arguments in favor of such measures seem to center on three

* In the U.S. Congress, Michael Harrington introduced a "truth in testing" bill in the last session of Congress, but the bill was never reported out of committee. In New York, in May 1977, eighteen state senators introduced a proposed amendment to the state education law which would require disclosure of test contents within thirty days of test administration. The proposal was opposed by the New York State Department of Education, the State University of New York at Albany, and the Educational Testing Service. With such opposition, it apparently stood little chance of enactment. However, in California on September 26, 1978 Governor Brown signed a bill which will require test sponsors to file copies of specified test questions and answers with the California Postsecondary Education Commission, to file certain information sufficient to describe the psychometric quality of the test, and to notify California test subjects regarding the purposes of the test, the subject matter of the test, the manner in which the test is scored, the manner in which the test scores will be recorded, the basis upon which scores will be made available, and a sample of test items. (For more information, see California Senate Bill No. 2005.)

themes: 1) test-takers should have the right to see corrected results as a simple matter of justice; 2) such freedom of information is necessary to help control the monopoly powers of test companies; and 3) prompt return of corrected results could be of real educational value, allowing test-takers to learn from their mistakes. Main arguments to the contrary are that: 1) such return of tests to test-takers would invalidate the tests for future use; 2) this would necessitate continual development of new tests for one-time administration, greatly increasing the costs and decreasing the quality of standardized testing; 3) lack of test security would exacerbate problems of test-coaching and teaching to the tests. There is not space here to consider such arguments in any detail. Nevertheless, it seems safe to speculate that such arguments, elaborating the tension implicit in the 1977 APA Ethical Standards between test-takers' rights and needs for test security will be increasingly common.*

Professional Standards and Public Practice. A major open question behind all of the foregoing discussion concerns the relationship between professional^{test} standards and public practice. There are at least two dimensions to this relationship--the match between the standards and the tests publishers develop, and the connection between the standards and test use. On both counts, the connections seem uncertain.

What is the relationship between test standards and test publishers'

* Indeed the issue has just come up in Massachusetts with regard to the state's policy for basic skills improvement. At a State Board of Education hearing on August 29, 1978, elementary school principal William Corbett proposed an amendment that state-required basic skills tests be returned to students after being corrected. The Board unanimously adopted the amendment, and Massachusetts State Department of Education officials are now puzzling over the amendment's implications.

practices? In a review of the 1966 version of the Standards for Educational and Psychological Tests, an official of the Educational Testing Service predicted that by the time his review appeared in print, "the major developers of tests will know the recommendations by heart and will be planning to include in future instruments those elements or standards which they have not already built into their publications" (Dobbins, 1966, pp. 752-53). Dobbins' prediction has been somewhat less than fully borne out. Consider, for example, the following report on the validity of the ETS-developed Sequential Tests of Educational Progress (STEP) Series II published in 1971--a full five years after the 1966 Standards:

The STEP Series II tests were designed to assess developed abilities in seven broad areas of education, and their content validity is of major importance. Content validity is best ensured by entrusting test construction to persons well-qualified to evaluate the relationship of the test items to current curricula and teaching objectives as was done for STEP Series II. (ETS, 1971, p. 150)

The STEP Series II manual presents a fairly general account of skill and content domains test items are designed to assess but completely fails to follow the 1966 Standards' requirement that in reporting on content validity, the test manual should indicate "clearly how adequate is the sampling" of the universe claimed to be tested (p. 15).*

Despite this example, there is evidence that some of the larger test publishers have in the past given attention to the test standards. In their study of the testing industry, for example, Holmen and Doctor (1972) observed

* It should be noted the STEP Series II manual is cited here as a case of non-adherence to the test standards not because it is a particularly bad example of a commercially published test manual. It is simply an example of minor irony, given Dobbins' prediction of 1966.

that

[The six largest] testing organizations provide models of comprehensive and technical service which represent very high standards of professional practice within the industry....With few exceptions, and most of these highly debatable, the [medium-sized testing] companies do a good job of meeting the requirements of the Test Standards Manual (pp. 42-43, 51).*

Oscar Buros (1974) has also observed that

There can be no question but that the Technical Recommendations and Standards have been effective in complementing the Mental Measurements Yearbooks in the drive for more honest and informative test manuals. Test manuals--as inadequate as most of them still are--would certainly have been even poorer had it not been for the Technical Recommendations, Standards, and the Mental Measurements Yearbooks (p. 757).

Buros nevertheless goes on to talk about the problem of nonadherence to the technical standards, and exclaims:

The nonobservance of the APA-AERA-NCME and MMY Standards by test authors, publishers, and users is shocking. When will test users make it unprofitable for publishers to market tests without reporting even the barest essentials of the data which were considered minimal forty-nine years ago? (Buros, 1974, p. 757).

More recent evidence of the weak impact of the test standards comes from a 1976 survey by the APA Committee on Psychological Tests and Assessments. The survey of the content of a number of major publishers'

* By "Test Standards Manual," Holmen and Doctor refer to the 1966 technical standards. In response to a query as to how the above conclusions were reached, Holmen (1978) explained that test publishers had made available "notes on the role of the standards in test design and validation--notes of specific test development projects, minutes of meetings discussing application of the standards, etc. We did not make test-by-test comparisons of the standards with the tests and their documentation except in a pilot study phase which convinced us the problems, with most of the companies, were not in failure to meet the standards, but with ignorant or careless use of tests."

test manuals "revealed that few manuals include all of the information indicated as essential in the test standards [1974]" (Singleton, n.d., p. 12). Also in 1977, Oles and Davis (1977) reported that test publishers generally seem to violate APA standards regarding test distribution.

The apparently small impact of the standards on test publishers practices may in part be a product of ambiguity in the standards themselves. To what extent do they represent levels of attainment regarded as measures of adequacy and to what extent do they represent standards of excellence. Generally the Standards seem to have opted for the latter with the result that test publishers have criticized them as "somewhat unrealistic, overly demanding, and impossible to meet without unwarranted investment" (Holmen and Doctor, 1972, p. 49).

Another explanation is the ambiguous status of test companies. On one hand they are part of a profession, presumably with allegiance to professional standards. On the other hand, however, they are commercial institutions guided by a very different set of standards. Writing in 1960, as to why the APA principle on test publication and advertisement "is violated more flagrantly than any other in our code of ethics," Wayne Holtzman observed "The enthusiastic author and obliging publisher cannot resist the temptation to sell the test as the latest word, the answer to eager personnel executives' problems" (Holtzman, 1960, p. 249, emphasis added). This ambiguous status was at the root of a 1967 case involving the Federal Trade Commission (FTC). At that time it was common practice for a number of larger test companies to exchange information concerning the qualifications of individuals who might wish to purchase tests. Such exchange made it easier for publishers to adhere to standards on distribution of tests (spelled out, as noted above in both the ethical and technical

standards). Such practices led to a complaint to the FTC, which subsequently charged four major test companies with restraint of trade. The case was resolved when the companies agreed to a consent decree to cease and desist from the practices charged (Holmen and Doctor, pp. 129-130). In other words, publishers' cooperation to implement professional standards ran afoul of standards concerning business practices. "Persons in the testing industry feel that, as a result of this complaint and the consent decree which followed it, tests were made available to persons not necessarily qualified to use them" (Holmen and Doctor, p. 130).

This other end of the standards-practice connection--namely, the link between the test standards and test use--also seems slight. The weakness of this connection has long been noted. For example, in a review of the 1966 Standards, Roger Lennon (1966) wrote:

There is little evidence to lead the reviewer to believe that many test users are now influenced in their selection of tests by the extent to which their manuals conform to the [1954] Technical Recommendations. Neither does there appear to be convincing evidence of correlation between improvement in test manuals and improvement in test use....[W]hat is called for is much better dissemination of the Standards, and better training of test users in the application of the information which, thanks to the Standards, is increasingly available in test manuals (p. 755).

Apparently, this recommendation met with little professional support, for none of the three organizations sponsoring the Standards (the APA, the AERA, and the NCME) seems to have taken strong action to disseminate the Standards or to promote their use.*

This situation seems odd, for as Harold Orlans has observed, the APA "has been the only major social science association which has taken effective steps to see that members observe ethical standards" (p. 80).

* The NCME is, however, now in the process of preparing a test users' guide to the 1974 joint Standards.

Indeed, the 1967 casebook on the 1963 Ethical Standards (APA, 1967) records eleven case in which the APA's Committee on Scientific and Professional Ethics and Conduct (CSPEC) took action to adjudicate apparent violations of the three 1963 ethical standards concerning testing. The following is an account of one of those cases:

A member of APA devised a test which was found by competent psychologists to be unreliable and invalid for the purposes for which claims were made. Responsible test publishers refused to publish or distribute the test, but the author had it printed privately and promoted its sale. He continued to do so after repeated advice and warning to desist, until finally he was dropped from membership by both the state association and APA.

Opinion. In evaluating tests, writings, or other professional productions of psychologists, care must be used to distinguish between a difference in point of view and failure to meet ethical standards. When, however, a number of competent peers conclude that the evidence indicates a test to be worthless, then commercial promotion of the test become unethical (APA, 1967, p. 58).

This example clearly indicates the limitations of this sort of professional review approach to the enforcement of testing standards. While the individual violating the standards was banished from the APA, there is no indication that the "unreliable and invalid" test was removed from the market.

In the past there have been several proposals for more stringent enforcement of testing standards. In 1950, an APA committee on test standards was charged with studying "the feasibility of and methods of implementing, a Bureau of Test Standards and a Seal of Approval to apply and enforce test standards," but took no action on these matters because of "the complex problems they presented" (Singleton, n.d., p. 3). More recently, a federally sponsored project on the Classification of Exceptional Children made a proposal along similar lines:

Because psychological tests of many kinds saturate our society and because their use can result in the irreversible deprivation of opportunity to many children, especially those already burdened by poverty and prejudice, we recommend that there be established a National Bureau of Standards for Psychological Tests and Testing [responsible to the Secretary of Health, Education, and Welfare]. (Hobbs, 1975, p. 237)

This proposal for federal intervention to curtail "poor tests and testing [which] may be as injurious to opportunity as impure food or drugs are injurious to health" (Hobbs, 1975, p. 238), apparently met with little enthusiasm.

So we are left in a curious situation. Much effort has gone into the development of professional standards concerning testing over a period of 25 years. But there is much evidence that the test standards have had limited direct impact on test publishers practices and even less on test use. Despite this apparent gap between standards and practice, there seems to have been little professional enthusiasm for concrete proposals to enforce standards on educational and psychological testing.

This raises the question of whether professional testing standards serve functions other than improving practice. Professionals seem reluctant to set up regular enforcement mechanisms for the enforcement of their standards in part because the notion of self-governance is part of the self-image of professionals. As Arlene Kaplan Daniels has observed, professional "codes do not simply fulfill the functions suggested by the professional ideology. Rather, they are part of the ideology, designed for public relations and justification for the status and prestige which professions assume vis-a-vis more lowly occupations"

Daniels, 1973, p. 49).*

This interpretation brings us back to the older meaning of the term "standard." Standard refers not just to criteria or measures of adequacy or excellence, but also to emblems or symbols of persons or groups.** In this meaning, professional standards have less to do with bettering professionals' practices than with burnishing professionals' public standing.

The public standing of professional groups increasingly seems in need of such support nowadays. Criticisms of professional groups' role in American society are fairly widespread recently (Bledstein, 1977; Larson, 1977), and the public seems more and more skeptical of the idea of professional

* Daniels quotes the ascerbic comments of a sociologist reacting to the idea of a code of ethics for his profession:

I want to register my protest against the latest effort to push through a code of ethics for sociologists. It seems to me that sociologists' own studies of occupational spheres, including the use of codes of ethics by various professional and business groups, should have convinced them that this is a fruitless undertaking. But apparently this is not the case. When we examine the codes of ethics of medicine, real estate, psychology, and a variety of other spheres of activities, we find that such codes are primarily designed to do the following:

1. Reduce the conflict among members of a professional of business fraternity (or at least the public manifestations of such conflict).
 2. Maintain a monopoly of lucrative professional tasks.
 3. Ward off criticism from outside the fraternity. Insofar as codes of ethics make a claim to protecting the public against improper professional or business practices, the codes are a fraud. The way in which they are actually used tends to protect the fraternity against the public rather than the other way around.
- (Quoted in Daniels, 1973, p. 49)

** In this respect it is of minor note that the only compilation of professional organizations' ethical codes which we could locate was contained in a volume entitled Professional Ethics and Insignia (Clapp, 1974).

self-governance in general. The apparent gap between testing standards and actual testing practice would seem to provide fuel for that skepticism. Whatever the public image of the testing profession, those who take seriously the idea of reforming testing practice with any sort of testing standards need at the outset to take a serious look at the connection between such standards and public practice.

#

REFERENCES

- APA (American Psychological Association). Ethical Standards of Psychologists. Washington, D. C.: APA, 1953.
- APA. Ethical Standards of Psychologists, American Psychologist, 14, pp. 279-282, 1959.
- APA. Casebook on Ethical Standards of Psychologists. Washington, D. C.: APA, 1967.
- APA. Ethical Standards of Psychologists. Washington, D. C.: APA, 1977.
- APA, AERA, & NCMUE. Technical Recommendations for Psychological Tests and Diagnostic Techniques (Reprinted in Buros, 1961). Washington, D. C.: APA, 1954.
- APA, AERA, & NCME. Standards for Educational and Psychological Tests and Manuals (Reprinted in Buros, 1972). Washington, D. C.: APA, 1966.
- APA, AERA, & NCME. Standards for Educational and Psychological Tests (Reprinted in Buros, 1974). Washington, D. C.: APA, 1974.
- APGA (American Personnel and Guidance Association). "Ethical Standards," Personnel and Guidance Journal, pp. 206-209, October 1961.
- APGA. Ethical Standards Casebook. Washington, D. C.: APGA, 1965.
- APGA. Ethical Standards. Washington, D. C.: APGA, 1972.
- APGA. Ethical Standards Casebook, Second Edition. Washington, D. C.: APGA, 1976.
- Bledstein, B. The Culture of Professionalism. New York: Norton, 1976.
- Buros, Oscar K. Tests in Print. Highland Park, N. J.: Gryphon, 1961.
- Buros, Oscar K. Personality Tests and Reviews. Highland Park, N. J.: Gryphon, 1972.
- Buros, Oscar K. Tests in Print II. Highland Park, N. J.: Gryphon, 1974.
- Clapp, J. Professional Ethics and Insignia. Metuchen, N. J.: Scarecrow Press, 1974.
- Cohen, David K. and Garet, Michael S. "Reforming Educational Policy with Applied Research," Harvard Educational Review, 45:1, February 1975.
- Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. The Dependability of Behavioral Measurements. New York: Wiley, 1972.

- Daniels, A. K. "How Free Should Professions Be?" in Freidson, E., ed. The Professions and Their Prospects, pp. 39-57. Beverly Hills, CA: Sage, 1973.
- Dobbins, J. E. "The Professional Service Representative of a Test Publisher" [Review of 1966 Standards for Educational and Psychological Tests and Manuals], Educational and Psychological Measurement, 26, pp. 751-753, 1966.
- ETS (Educational Testing Service). Handbook for STEP Series II. Princeton, N. J.: ETS, 1971.
- Goldstein, H. et al. "Schools," in Hobbs, N. ed., Issues in the Classification of Children, Volume II, pp. 4-61. San Francisco: Jossey-Bass, 1975.
- Hobbs, N. ed. The Futures of Children: Report of the Project on Classification of Exceptional Children. San Francisco: Jossey-Bass, 1975.
- Holmen, M. G. Personal communication to Walter Haney of The Huron Institute, September 1978.
- Holmen, M. G. and Doctor, R. Educational and Psychological Testing: A Study of the Industry and its Practices. New York: Sage, 1972.
- Holtzman, W. "Some Problems of Defining Ethical Behavior," American Psychologist, 15:3, pp. 247-250, March 1960.
- Kelley, T. L. Statistical Method. New York: MacMillan, 1924.
- Larson, M. S. The Rise of Professionalism: A Sociological Analysis. Berkeley, CA: University of California Press, 1977.
- Lennon, R. "A Measurement Specialist and Executive of a Large Test Publishing Firm" [Review of 1966 Standards for Educational and Psychological Tests and Manuals], Educational and Psychological Measurement, 26, pp. 753-756, 1966.
- Lerner, B. "The Supreme Court and the APA, AERA, NCME Test Standards: Past References and Future Possibilities," Draft paper presented at a Conference of the National Consortium on Testing, May 15, 1978.
- Oles, H. J. and Davis, G. D. "Publishers Violate APA Standards on Test Distribution," Psychological Reports, 41, pp. 713-714, 1977.
- Orlans, H. Contracting for Knowledge. San Francisco: Jossey-Bass, 1973.
- Singleton, M. C. "Historical Survey of Issues Considered by the Committee on Psychological Tests and Assessment: 1895-1976," unpublished paper, APA, n.d.
- Stanley, J. C. "Reliability," in Thorndike, R. L., ed. Educational Measurement, Second Edition, pp. 356-442. Washington, D. C.: American Council on Education, 1971.